We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

172,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Ontological description of gene groups by the multi-attribute statistically significant logical rules

Aleksandra Gruca and Marek Sikora
Silesian University of Technology, Institute of Informatics
Poland

1. Introduction

With the beginning of the post-genomic era, the DNA microarray chips became one of the most indispensable tools used in the biological laboratories all over the world to support the genome scale experiments (Baldi & Hatfield, 2002). Development of the DNA chip technologies provided a great opportunity to learn and understand biological phenomena of the world of living organisms better by identifying the genes involved in various biological processes.

A single DNA microarray experiment allows to record simultaneously thousands of gene expression profiles obtained under similar experimental conditions. Such abundance of data needs to be analyzed and processed by means of sophisticated mathematical tools. Nowadays it is impossible for a single person to understand and interpret the experiment results which consist of thousands of numbers describing the expression profiles of all the genes whose probes are located on the DNA microarray chip. Only the cooperation among biologists, mathematicians and computer scientists leads to the complete and profound understanding of the biological phenomena that are recorded in the laboratory.

The analysis of the data obtained in the DNA microarray experiment is a complex and difficult process involving applications of many specialized algorithms. Various methods such as statistical analysis and data mining techniques need to be applied until a final biological conclusions are derived from the spot intensity measurements obtained from a scanner software.

Such complex analysis usually consists of the four main steps: data normalization, identification of the differentially expressed genes, application of the algorithms grouping together genes with similar expression patterns and interpretation of the biological functions of the genes co-expressed together.

Data normalization involves methods for reducing effect of the systematic errors. Differentially expressed genes are identified by applying statistical tests. Unsupervised methods organize genes expression into separate sets (clusters) that co-expressed together, thus the groups obtained are expected to include genes involved into similar biological processes.

One of the most important tasks is the interpretation of the obtained clusters – genes that form the clusters need to be confronted with existing biological knowledge on its classes and functions. The knowledge of genes and their functions included in the analysis supports extracting biological knowledge from the performed experiments. This aspect of data analysis is often done by an expert in the field of the experimental design, frequently manually, what is time consuming for large data sets.

However, the expert work may be supported by specialized systems designed for storing, organizing and extracting the relevant information and the new algorithms incorporating the existing knowledge about genes. These methods, including knowledge discovery, artificial intelligence and automatic reasoning appeared to be very useful and efficient on the field of analyses of specific biological data.

2. Gene Ontology analysis

One of the most popular and widely used sources of information about biological processes and genes involved in these processes is the Gene Ontology database (Ashburner et al., 2002). Gene Ontology consortium provides structured and controlled vocabulary that is used to describe genes and their products independently of the species. The GO database is organized into three disjoint directed-acyclic graphs (DAGs) describing biological process (BP), molecular function (MF) and cellular component (CC). Each node of the graph is called a GO term, and it is a single unit that describes some known biological process or function of the gene. The dependences between GO terms are hierarchical and as the DAG is traversed from the root into its leafs, the terms are inspected from the general ones to the more specific concepts. Figure 1 presents part of the GO directed-acyclic graph structure.

Fig. 1. Gene Ontology directed-acyclic graph structure

Standard approach to the analysis with the use of the GO terms includes performing various statistical tests to detect enrichment or depletion of the GO terms in the analyzed gene group. There is a number of GO processing tools available, based on the same idea of mapping GO terms on the genes composing the cluster and then performing a statistical test to extract over- or underrepresented GO terms in the analyzed set of genes. Thus, a Bingo (Maere et al., 2005), Cytoscape plugin, implements a hypergeometric and binominal test for the over-representation of the GO terms and then visualizes the results by means of Cytoscape network (Cline et al., 2007). Another example of the GO processing tool is a FatiGO (Al-Shahrour et al., 2005), a web-based application which allows to identify over- or underrepresented GO terms by applying Fisher's exact test and performs the analysis separately for each GO level. Most of the tools consider using multiple-testing corrections what reduces the rate of false positives occurring among significant results when testing of thousands hypothesis in the same experiment (Dudoit et al., 2003).

GO browsers such as Bingo or FatiGO usually take as an input two lists of genes: one of them represents investigated gene set, the other – rest of the genes involved in the microarray experiment. Each gene from the both lists is then annotated with the appropriate GO terms and the statistical test is applied to check for significant over or under-representation of GO terms in one of the sets with respect to another one. Thus, the final result obtained from such analysis is the list of the statistically significant GO terms that are enriched or depleted in the investigated gene set. For each term, the p-value of the statistical test corresponding to that term is recorded and by comparing the p-values of all GO terms describing genes belonging to the cluster one can have a good indication about the prominence of a given functional GO category in the gene set.

3. Related work

Knowledge based approaches were used in Gene Ontology analysis mainly for genes grouping with respect to their expression level (Eissen et al., 1998, Iyer et al., 1999) and describing groups of similar genes by means of a set of Gene Ontology terms (Brown et al. 2002; Khatri & Draghici, 2005).

The research strictly connected with Gene Ontology analysis by means of rule induction algorithms were conducted by Hvidstein (Hvidstein et al., 2003) and Midelfart (Midelfart, 2005a, Midelfart, 2005b). In the paper (Hvidstein et al., 2003) conditional rules of the form *"IF conjunction of conditions describing time series of gene expression profile THEN ontological term"* were created. The authors wanted to assign genes with specified expression profiles to a specific gene ontology term. To simplify the analysis process and get more general rules, a set of gene ontology terms was limited in such manner that only terms located on the specified ontology level were considered during the analysis. The ROSETTA (Øhrn et al., 1998) software based on rough sets methodology (Pawlak, 1991) is used for induction of rules. Conclusions of rules with the same conditional parts are joined, thus the rules describing a group of genes with similar expression profiles are obtained. In the conclusions of the rules, the set of gene ontology terms describing the group is included. The research shows that number of the obtained rules is huge and there is a need to apply a rule filtration method (Sikora, 2006) or pointing to a user so-called the most interesting rules (Agrawal & Srikant, 1994).

In the Midelfart paper a similar approach was used, except that there was not defined which terms can be located in the rule conclusions. The induction algorithm considers the fact that terms (what means also rule conclusion) are related (according to the ontology graph). Beginning from the terms lying at the lowest level of the ontology graph, the algorithm generates rules for each term. If the rules generated are recognized as good quality rules, they are added into a result set of rules, and all genes recognizing the rules are removed from the analyzed data set. If the rules describing a given ontology term are not of good quality, they are removed from the description and genes assigned to the analyzed term are moved to higher ontology level, and learning process proceeds again for a new (placed higher) term. Accuracy and generality decide about a rule quality and thresholds values of the both parameters have to be defined by a user. From the experiments performed follows that the method given by Midelfart allows obtaining better results than the method proposed by Hvdistein. Both papers verify quality of obtained results by means of a classification accuracy, that is by rightness of assignment to corresponding GO terms these genes that were not taken into consideration during the analysis. Midelfart also introduces a measure that presents how many ontology terms managed to describe (assign) genes placed in the analyzed data set.

Both of the methods presented above do not allow obtaining an information about influence of co-appearing of some ontology terms on a gene membership to a specified gene group. The methods do not consider statistical aspect of determined rules, that is they do not search whether discovered dependences are statistically significant (Rice, 1995).

None of the published papers include research on describing already defined group of similar genes by means of multiattributes logical rules. Conjunctions of ontology terms are placed in premises of rules and conclusion include a specific group of similar genes - research on one-attribute rules were conducted only (Gruca, 2008).

However, the similar research has been conducted recently by Carmona-Saez and coworkers (Carmona-Saez et al., 2007). The authors propose a method Genecodis, a web-based tool for integrated analysis of annotations from different sources. They applied the classical Apriori algorithm (Agrawal & Srikant, 1994) to discover association rules to extract a sets of annotations (so-called itemsets) that frequently co-occur in analyzed group of genes. In order to reduce the number of possible outcome combinations, and thus the computation time, the authors introduced the threshold, so-called support value, that allows discovering only such combinations of annotations that appear in at least x genes. The algorithm starts by determining a set of single annotations that satisfy the minimal support condition x . Then, combinations of annotations are created on the basis of the set of single annotations obtained in a first step of algorithm. Each combination generated must describe at least x genes. The combinations obtained are further analyzed in order to verify statistical significance of computed concurrencies.

On the basis of this method Hackenberg (Hackenberg & Matthiesen, 2008) developed another web-based application that uses various annotation databases and that allows a user to discover combinations of annotations from many different fields such as functional categories, gene regulation, sequence properties, evolution, conservation, etc.

Association rules are also considered in (Carmona-Saez et al., 2006). The method proposed in this paper combines expression data and biological information. Extracted association rules are of the following form: $\{cell\ cycle \rightarrow [+]\ condition\ 1, [+]\ condition2, [-]\ condition3\}$ which means that in the data set a significant number of genes annotated as "cell cycle" are over-

expressed in condition 1 and 2 and under-expressed in condition 3 (Carmona-Saez et al., 2006). Several measures such as support, confidence and improvement are applied to compute the significance of the rule.

In our research we assume that there are the groups of genes with similar expression profiles. The groups were obtained by expert analysis, or by means of clustering method (Eisen et al., 1998; Kustra & Zagdański, 2006). The aim of this paper is to present a method that allow to describe each of obtained group of genes with the use of gene ontology terms. The rules determined have the following meaning:

“if a gene is described by a conjunction of gene ontology terms appearing in a rule premise, then it belongs to a specific group of genes”.

Covering all genes belonging to a given group we obtain a set of rules which can be interpreted as follows:

“a given group of genes is described by gene ontology terms appearing in at least one rule indicating the described group”.

It is also important that the rules generated have a statistical significance level which can be determined by a user. Obtained rules also consider co-occurrence of the terms in a given gene group and the presented method guarantees that the co-occurrence will not be trivial (for example, resulting from hierarchy of the ontology graph).

4. Description of gene groups by multi-attribute logical rules

4.1 Basic notions

Let there be a set of genes G and a set of descriptions of genes and gene products A . There is a hierarchical relationship among elements from the set A in a form of directed-acyclic graph (DAG) (Fig.1). This graph is interpreted as a database, called Gene Ontology. Each element of the Gene Ontology database is called GO term which is represented by a name and a unique 7-digits identifier. There are three main ontologies describing genes and gene products from three independent biological domains: biological process, molecular function and cellular component. Each of them is represented as a separate DAG. The root of each graph is also the name of the biological domain that is described by that ontology.

Formally, Gene Ontology is a directed acyclic graph $GO=(A,\leq)$, where A is a set of GO terms describing genes and its products and $\leq$ is a binary relation on A such that genes described by GO term a_j are a subset of genes described by GO term a_i , denoted $a_j \leq a_i$, if and only if there exist a path $(a_i, a_{i+1}, \dots, a_{j-1}, a_j)$ such that $a_m \leq a_{m-1}$ for $m=i+1, i+2, \dots, j-1, j$ (we can find here an analogy to the inverse Hasse diagram for ordering relations). Relation $\leq$ is the order relation (reflexive, antisymmetric and transitive). The root of the DAG is the largest element and we assume that the root is on a level zero in the ontology. Each level of the graph is defined in the following way: the i -th level of the graph is formed by all GO terms $a \in A$ for which a path exists $(root, a_1, \dots, a_{i-1}, a_i)$ such that: $a_1 \leq root$, $a_m \leq a_{m-1}$ for $m=2, 3, \dots, i-1$ and $a_i \leq a_{i-1}$ (in other words there exists a path of length i from the root to that GO term). The hierarchical structure of the ontology database allows representing biological knowledge on the multiple levels of details. Terms at the higher levels (closer to the root) describe more general

function or process while terms at the lower levels are more specific. Each node from the DAG is represented by a single GO term from the set A . These terms are further applied in the annotations of genes or gene products in the biological databases. Each annotation is an association between gene and the GO term describing it, thus for a simplification, we can assume that each node of the Gene Ontology is also annotated by the genes from the set G . A gene can be annotated to zero or more nodes for each ontology, at any level within each ontology. All GO terms that exist in the DAG must follow the true path rule: "the pathway from a child term all the way up to its top-level parent(s) must always be true". The consequence of such approach is that annotating gene to a GO term implies annotation to all parents via any path.

It stems from the definition of the Gene Ontology and from the interpretation of the relation $\leq$ that each gene annotated with GO term $a_j \in A$ is also annotated with a GO term $a_i \in A$ such that $a_j \leq a_i$. The relation $\leq$ is transitive, so if we take into consideration the hierarchical structure of the DAG we may assume that each gene annotated with the particular GO term (that was assigned to that gene by a curator from Gene Ontology consortium), is also annotated with the GO terms that are parents of that particular term. In the other words – the higher is the level of a GO term, the more genes are annotated to that term. Assuming that G_a is a set of genes annotated with the GO term corresponding to the node a , for each node a_i such that $a \leq a_i$, $G_a \subseteq G_{a_i}$ is satisfied.

To preserve the clarity of the ontology, the annotation files that are available at the Gene Ontology consortium website include only "original" annotations, that is annotations that were assigned to the particular GO terms by a curator. The annotations resulting from the "true path rule" (annotation of that gene to all parent nodes of that term) are not included in the annotation files. In our analysis we consider a graph that is constructed as described above. To increase the number of genes annotated to each node (GO term) of the graph, we assign to each node not only the genes that were directly extracted from the Gene Ontology annotation database but also the genes that are annotated to all descendant terms of that node. We call such a graph GO-Inc (similar approach called inclusive analysis is introduced in GO browser FatiGO).

To summarize, there are given: a set G of genes, a set A of GO terms that create GO-Inc ontology graph and n gene groups with similar expression profiles $\{G(1), G(2), \dots, G(n)\}$. It is possible to create a decision table $DT = (G, A \cup \{d\})$, where for all $a \in A$, $a: G \rightarrow \{0, 1\}$, and $d(g) \in \{G(1), G(2), \dots, G(n)\}$ for all $g \in G$. Thus, rows in the table DT contain descriptions of the single genes belonging to the set G created by means of GO terms from A . The notation $a(g)=1$ denotes that the gene g is assigned to the term a in the GO-Inc graph. Beside description by the terms from A , each gene is also characterized by a membership to specific group of similar genes (value $d(g)$). In the table DT we try to find all statistically significant relationships in the rules of the form (1)

$$\text{IF } a_{i1}=1 \text{ and } a_{i2}=1 \text{ and } \dots \text{ and } a_{ik}=1 \text{ THEN } d=G(l) \quad (1)$$

or in the shortened notation:

$$\text{IF } a_{i1} \text{ and } a_{i2} \text{ and } \dots \text{ and } a_{ik} \text{ THEN } d=G(l) \quad (2)$$

where: $\{a_{i1}, a_{i2}, \dots, a_{ik}\} \subseteq A$, $G(l) \in \{G(1), G(2), \dots, G(n)\}$.

A rule of the form (2) should be interpreted as follows:

*IF a gene is described simultaneously by the terms occurring in a premise of the rule
THEN it belongs to the gene group $G(l)$*

The above form of the rule representation denotes that descriptors $a_{i1}=0$ are not in our interest now, that is we do not describe gene group by means of the statements “GO term a does not describe a given gene group”.

A set of rules with identical conclusions we denote by $RUL_{G(l)}$ and call the description of the gene group $G(l)$.

4.2 Rules quality measures

Rules of the form (2) are the special case of so-called decision rules, and several measures that reflect quality of a decision rule can be connected with the rule. These measures are called the rule quality measures and their main purpose is a rule evaluation and/or such steering of induction and/or reduction processes that there are rules of the best quality in the output rule set. Values of most known rule quality measures (An & Cercone, 2001; Sikora, 2006) can be determined based on the analysis of a contingency table, that allows describing rules behavior with relation to the decision table DT .

A set of genes which are described by the terms occurring in a premise of a rule r of the form (2) we denote by $match(r)$. A set of genes characterized by the terms occurring in the premise of r and belonging to a group indicated by the rule conclusion we denote by $supp(r)$. If the rule r is shortly written as $\varphi \rightarrow \psi$, then the contingency table of the rule has the following form:

$n_{\varphi\psi}$	$n_{\varphi\neg\psi}$	n_{φ}
$n_{\neg\varphi\psi}$	$n_{\neg\varphi\neg\psi}$	$n_{\neg\varphi}$
n_{ψ}	$n_{\neg\psi}$	n

Table 1. Contingency table of the rule $r \equiv \varphi \rightarrow \psi$

In Table 1, $n_{\varphi} = n_{\varphi\psi} + n_{\varphi\neg\psi} = |G_{\varphi}|$ is the number of genes that recognize the rule $\varphi \rightarrow \psi$; $n_{\neg\varphi} = n_{\neg\varphi\psi} + n_{\neg\varphi\neg\psi} = |G_{\neg\varphi}|$ is the number of genes that do not recognize the rule $\varphi \rightarrow \psi$; $n_{\psi} = n_{\varphi\psi} + n_{\neg\varphi\psi} = |G_{\psi}|$ is the number of genes that belong to the gene group described by the rule $\varphi \rightarrow \psi$; $n_{\neg\psi} = n_{\varphi\neg\psi} + n_{\neg\varphi\neg\psi} = |G_{\neg\psi}|$ is the number of genes that do not belong to the gene group described by the rule $\varphi \rightarrow \psi$; $n_{\varphi\psi} = |G_{\varphi} \cap G_{\psi}|$ is the number of genes that support the rule $\varphi \rightarrow \psi$; $n_{\varphi\neg\psi} = |G_{\varphi} \cap G_{\neg\psi}|$; $n_{\neg\varphi\psi} = |G_{\neg\varphi} \cap G_{\psi}|$; $n_{\neg\varphi\neg\psi} = |G_{\neg\varphi} \cap G_{\neg\psi}|$.

Using the information included in the contingency table and the fact that for a known rule $\varphi \rightarrow \psi$, there are known the values $|G_{\psi}|$ and $|G_{\neg\psi}|$, it is possible to determine the values of quality measures based on the values $n_{\varphi\psi}$ and $n_{\varphi\neg\psi}$. It can be also noticed that for any rule $\varphi \rightarrow \psi$ the inequalities $1 \leq n_{\varphi\psi} \leq |G_{\psi}|$, $0 \leq n_{\varphi\neg\psi} \leq |G_{\neg\psi}|$ hold. Hence, the quality measure is the function of two variables $q(\varphi \rightarrow \psi): \{1, \dots, |G_{\psi}|\} \times \{0, \dots, |G_{\neg\psi}|\} \rightarrow \mathbf{R}$.

Two basic quality measures are accuracy (3) and coverage (4) of a rule. Both the measures considered at the same time provide complete objective (referring to DT) evaluation of the rule quality, since according to the principle of induction by enumeration it is known that the rules with a good accuracy and coverage reflect real dependences. The dependences are true also for the objects from outside of the analyzed dataset. It is easy to prove that along

with accuracy increasing, rule coverage decreases. Therefore, attempts are performed to define objective quality measures that respect accuracy and coverage of a rule simultaneously (Bruha, 1997; An & Cercone, 2001; Furnkranz & Flach, 2005 ; Sikora, 2006). In the researches quoted some evaluation measures achieved the good results both in a classification accuracy (generalization ability) and a size of classifiers (description ability).

$$\text{acc}(r) = \frac{n_{\varphi\psi}}{n_{\varphi}} \quad (3)$$

$$\text{cov}(r) = \frac{n_{\varphi\psi}}{n_{\psi}} \quad (4)$$

In our research, a rule quality measure is used in filtration process in order to limit a number of determined rules. We applied a modified version of the WS (5) measure proposed by Michalski (Michalski, 1983).

$$\text{WS}(r) = w_1 \text{acc}(r) + w_2 \text{cov}(r) \quad (5)$$

In a rule induction system YAILS values of parameters w_1, w_2 for the rule r are calculated as follows: $w_1 = 0.5 + 0.25 \text{acc}(r)$, $w_2 = 0.5 - 0.25 \text{acc}(r)$. The WS measure is monotone with respect to each variable $n_{\varphi\psi}$ and $n_{\varphi \rightarrow \psi}$, and assumes values from the interval $[0,1]$ (Fig.2).

Fig. 2. Graph of WS-Yails measure for the rule $\varphi \rightarrow \psi$ depending on $n_{\varphi\psi} \equiv \text{supp}(\varphi \rightarrow \psi)$ and $n_{\varphi \rightarrow \psi} \equiv \text{match}(\varphi \rightarrow \psi) - \text{supp}(\varphi \rightarrow \psi)$. A number of genes described by the rule is $n_{\psi} = 25$, and a number of remaining genes is $n_{\neg\psi} = 250$.

We assume, that the rules determined of the form (2), have to be statistically significant rules, with a significance level better then or equal to a level established by a user. To evaluate a statistical significance of created rules we use a hypergeometric test (6). It can be easily noticed that a p-value for the rule r of the form (2) can be determined on the basis of the contingency table (Table 1) and computed using the formula (7). P-values obtained for fixed n_{ψ} and $n_{\neg\psi}$ are presented on a graph in Fig.3. To verify the statistical significance of the

rules induced, we also compute false discovery rate (FDR) coefficient (Benjamini & Hochberg, 1995)

$$p(n_{\varphi\psi}, n_{\varphi\neg\psi}, n_{\neg\varphi\psi}, n_{\neg\varphi\neg\psi}) = \frac{\binom{n_{\varphi}}{n_{\varphi\psi}} \binom{n_{\neg\varphi}}{n_{\varphi\neg\psi}}}{\binom{n}{n_{\psi}}} \quad (6)$$

$$p_{val}(n_{\varphi\psi}, n_{\varphi\neg\psi}, n_{\neg\varphi\psi}, n_{\neg\varphi\neg\psi}) = \sum_{k=0}^{n_{\varphi\neg\psi}} p(n_{\varphi\psi} + k, n_{\varphi\neg\psi} - k, n_{\neg\varphi\psi}, n_{\neg\varphi\neg\psi}) \quad (7)$$

Fig. 3. Graph of p-values depending on $n_{\varphi\psi} \equiv \text{supp}(\varphi \rightarrow \psi)$ and $n_{\varphi\neg\psi} \equiv \text{match}(\varphi \rightarrow \psi) - \text{supp}(\varphi \rightarrow \psi)$. A number of genes described by the rule is $n_{\psi} = 10$, a number of remaining genes is $n_{\neg\psi} = 250$.

4.3 Rules induction

Induction of decision rules can be classification or discovery oriented. The purpose of classification-oriented induction is to find, on the basis of the set of learning examples, a set of decision rules that will be used to classify new unknown examples (Michalski et al., 1998; Mitchell, 1997). The purpose of discovery-oriented induction is to discover patterns and regularities (in the form of IF THEN rules) in data which are interesting and useful to different kinds of users. A set of classification rules often contains only a limited part of all possible rules that could be induced from that set. Majority of the classification-oriented algorithms follow the same general scheme using a greedy heuristic strategy. This scheme consists of creating a first rule by choosing sequentially the 'best' elementary conditions according to some quality criteria. Then, learning examples that match this rule are removed from the set considered. The process is repeated iteratively until all the examples are covered by the result rule set. An alternative approach, better from the knowledge discovery point of view, is the induction of all the rules that satisfy some requirements. This approach is implemented among others in the association rules induction algorithm Apriori (Agrawal & Srikant, 1994), and the decision rules induction algorithm Explore (Stefanowski & Vanderpooten, 2001). Another possibility is to induce all rules and then filter them to find the most interesting ones (Klemettinen, 1994; Sikora, 2009).

We generate the rules with p-values less or equal to a threshold established by a user. Therefore to describe given gene group, we must determine all possible combinations of all possible subsets of GO terms. Since we are interested in the premises with descriptors assuming value one (see rule definition (2)), in a pessimistic case we have to determine

$$\sum_{k=1}^{|A|} \binom{|A|}{k} = 2^{|A|} - 1 \text{ rules, what is impossible in the case of a big number of considered GO}$$

terms. We can model on the heuristic algorithm Apriori used for association rules searching during rules determining. In contrast to association rules, in our case a rule conclusion is already established. Therefore, Explore algorithm proposed by Stefanowski is more suitable for our aims. Using Explore algorithm we can generate all possible conjunctions of GO terms for each gene group. To limit a number of analyzed combinations, a user can provide minimal requirements concerning the quality of determined rules (for example, minimal rule coverage, minimal rule accuracy). In the method proposed, we generate only statistically significant conjunctions of GO terms.

In the standard version of Explore algorithm, rule induction starts from one-condition rules and then next conditions are iteratively added to each rule. After adding a next condition, a created rule is verified whether it satisfies quality criteria defined by a user. If it does, then the rule is added to a result rule set. If it does not, then the algorithm checks whether the created (temporary) rule may satisfy quality criteria introduced by the user (for example, whether the rule of the current form covers enough quantity of objects – minimal coverage requirements). A premise of the temporal rule that has a chance of satisfying minimal requirements of quality is widen by the next conditions, other temporal rules are not widen and are removed. A detailed description of the algorithm can be found in (Stefanowski & Vanderpooten, 2001).

For our purposes, Explore algorithm was put to a few modifications. Searching space of potential candidates for rules is made by means of a procedure that is iteratively repeated for each gene group. The main part of the algorithm generates premises with increasing size, beginning from premises containing one GO term. When a rule created in this way satisfies a p-value criterion established by a user, it is added to the generated rule set and conjunction is widen (assuming that other statistically significant rules can be determined from the conjunction). Motivation of such solution is that more information may be obtained from the statistically significant rule (8) than also significant, but shorter rule (9).

IF primary metabolic process **and**
 membrane-bounded organelle **and**
 proteasome storage granule **and**
 threonine-type endopeptidase activity **and**
 proteasome core complex
THEN gene group 2
 supp(r)=9, match(r)=9,
 acc(r)=1, cov(r)=0.33, pVal(r)=2.231392e-010

(8)

IF primary metabolic process **and**
 membrane-bounded organelle
THEN gene group 2
 supp(r)=19, match(r)=114,
 acc(r)=0.166, cov(r)=0.703, pVal(r)=0.00147

(9)

If for a given premise all GO terms were already considered, then a new GO term is selected (not chosen yet) and a new rule creation begins.

In order to narrow the searching space the following solutions were applied:

- After adding a GO term a to a rule premise, no terms lying on any path (from ontology leaf to the root) that leads to the term a are considered. Let us notice that for any term $b \in A$ for which $b \leq a$, the conjunction $b \wedge a$ reduces to a , since b lies on lower level than a , so $G_b \subseteq G_a$. Similarly, for $c \in A$ for which $a \leq c$, the conjunction $a \wedge c$ reduces to c , since c lies on a higher level. There is no point to consider the conjunction $a \wedge c$ because the descriptor c will be considered by the algorithm separately. The same reasoning can be presented for a conjunction composed of bigger number of terms to which a new GO term is added.
- If a rule with the premise $a_1 \wedge a_2 \wedge \dots \wedge a_j$ was added to the result set of rules, then rules with premises being permutations of $\{a_1, a_2, \dots, a_j\}$ will be no longer considered. The algorithm starts creation of a rule premise by selecting the terms that describe the most genes in the analyzed gene group.
- Assuming that the currently created rule has the form $\varphi \rightarrow \psi$, the GO term a will be added to its premise forming the rule $\varphi \wedge a \rightarrow \psi$, if $\text{acc}(\varphi \rightarrow \psi) \leq \text{acc}(\varphi \wedge a \rightarrow \psi)$. The condition limits a number of analyzed combinations by rejecting the terms that do not contribute to improving a rule accuracy, thus improving statistical significance of the rule as well.

The algorithm starts with a parameter characterizing maximal number of terms in the rule premises. However, a number of rules determined in this way is still very large. Therefore, a method of rules evaluation and filtration is required.

4.4 Rules evaluation and filtration

Statistical significance of a rule is the first criterion of its evaluation. A graph of p-value depending on a number of objects recognizing and supporting the rule is presented on the Fig. 3. From the graph follows that evaluating statistical significance only may lead in some cases to prefer hardly accurate rules. A rule describing one of gene group from the YEAST set (Eisen et al., 1998) is such example (10):

$$\begin{aligned} &\text{IF cellular metabolic process and} \\ &\text{macromolecule metabolic process} \\ &\text{THEN gene group 2} \\ &\text{acc}(r)=0.116, \text{cov}(r)=1, \text{pVal}=0.0077, \text{FDR}(r)=0.00872 \end{aligned} \quad (10)$$

The group consists of 27 genes, and there are 274 genes in the analyzed set. The presented rule covers the whole group, and recognizes 231 genes. Thus it is very inaccurate one. The rule is low-quality rule, however, significance level of the rule is good. This follows from the fact that the rule concentrates probability of selecting genes from the group described by the rule with relation to whole data set. The presented example shows that we need other rule quality evaluation criteria in order to select the best rules from the analyzed set of rules. Thus, the next criterion evaluating rule quality is a modified measure WS (11):

$$\text{mWS}(r)=[0.5+0.25\text{acc}(r)]\text{acc}(r)+[0.5-0.25\text{cov}(r)]\text{cov}(r) \quad (11)$$

It can be easily noticed that the measure mWS differs from the measure applied in the system YAILS (5) by a manner of calculation a value of the parameter w_2 . In the modification presented in the formula (11) the rule coverage was used for establishing the value of the parameter w_2 unlike the rule accuracy used for the measure (5). The change is justified by the fact that for general, inaccurate rules the modified version of WS (11) assumes less values than original WS measure. Differences in rule evaluation between measures (5) and (11) can be clearly seen by comparing graphs of the both measures (Fig. 4).

Fig. 4. Graphs of the standard and modified WS Yails measure

Another criterion of a rule quality evaluation is a number of GO terms included in a rule premise (12). We assume that the bigger number of terms is in the rule premise, the more information is represented by rule (we remind that terms occurring in a premise do not lie on a common path in ontology graph).

$$\text{length}(r) = \frac{\text{NoGOterms}(r)}{\text{MaxGOterms}} \quad (12)$$

where $\text{NoGOterms}(r)$ is the number of GO terms in the r rule premise, MaxGOterms is the maximal number of descriptors in the longest rule that describes the same gene group as the rule r .

The last rule quality evaluation criterion is the level of GO terms occurring in the rule premise (13). From a description point of view we should prefer rules with premises including terms from as low level of the GO graph as possible.

$$\text{depth}(r) = \frac{\sum_{i=1}^{\text{NoGOterms}(r)} \text{level}(a_i)}{\sum_{i=1}^{\text{NoGOterms}(r)} \text{max_path}(a_i)} \quad (13)$$

where: $\text{level}(a_i)$ is the level of GO term a_i that occurs in the rule premise, $\text{max_path}(a_i)$ is the longest path leading from the root to leaf of GO-Inc that pass through the node a_i .

Finally, a measure that incorporates all aspects of rule quality evaluation presented above, is a product of all component measures (14):

$$Q(r) = mWS(r) \times \text{length}(r) \times \text{depth}(r) \quad (14)$$

The measure Q is monotone with respect to each component of the product, i.e. when values of any two factors are fixed, the measure value is non-decreasing function with respect to the third factor. This feature fulfills assumptions of quality measures that were given by Piatetsky-Shapiro (Piatetsky-Shapiro, 1991).

The compound quality measure Q allows creating a ranking of rules describing each gene group. Below we present a filtration algorithm that uses obtained rules ranking:

Rules filtration algorithm

Input: G – a genes set, RUL_G a set of rules describing the gene group G .

Output: RUL_G^F – filtered rule set covering the same objects from the set G as the set RUL_G .

Begin

$RUL_G^F := \emptyset$;

Create a ranking $RANK_RUL$ of rules from the set RUL_G according to decreasing value of an evaluation measure Q ;

While $RANK_RUL \neq \emptyset$ **or** $G \neq \emptyset$ **do**

Begin

Take a first rule r from the set $RANK_RUL$;

$RANK_RUL := RANK_RUL - \{r\}$;

$RUL_G^F := RUL_G^F \cup \{r\}$;

For all $rr \in RANK_RUL$ **do**

Begin

If $\text{supp}(rr) \subseteq \text{supp}(r)$ **then**

Begin

If $\text{sim}(rr, r) \leq \varepsilon$ **then** $RUL_G^F := RUL_G^F \cup \{rr\}$;

$RANK_RUL := RANK_RUL - \{rr\}$

End if

End for

Renumber the set $RANK_RUL$ taking into account removed rules;

$G := G - \text{supp}(r)$;

End while

End

The presented filtration algorithm creates a coverage of the set G (or part of the set G which is covered by rules from the set RUL_G) beginning from the best rules with respect to the value of measure Q . If the rule rr covers the same objects as currently the best rule r , but contains other biological knowledge, then rule rr is not removed but remains in the result set of rules. It is worth to notice that the filtration algorithm defined in such way does not guarantee that filtered rule set RUL_G^F includes only dissimilar rules containing different biological knowledge. If we assume that there are three initial rules in the ranking r_1, r_2, r_3 such as: $\text{supp}(r_3) \subseteq \text{supp}(r_1)$, $\text{supp}(r_2) \subseteq \text{supp}(r_1)$ and $\text{sim}(r_1, r_3) \leq \varepsilon$, $\text{sim}(r_1, r_2) \leq \varepsilon$, then the rules

r_2, r_3 will be included in the result data set. Meanwhile, it may appear that $\text{sim}(r_2, r_3) > \varepsilon$ which means that rules r_2, r_3 include similar biological knowledge. However, the presented filtration algorithm do not remove neither the rule r_2 nor r_3 . Therefore, after the first part of the filtration, the second part consist in verification of rules similarity in pairs. If rules are similar, then a rule covering more objects always remains in the filtered rules set.

The similarity of rules is determined by the formula (15). If a rule is similar to the reference rule to degree greater than 50% ($\varepsilon=0.5$), then it is finally removed from the set of determined rules, otherwise it remains in an output rule set.

$$\text{sim}(r_i, r_j) = \frac{\text{uGOterms}(r_i, r_j) + \text{uGOterms}(r_j, r_i)}{\text{NoGOterms}(r_i) + \text{NoGOterms}(r_j)} \quad (15)$$

where: $\text{uGOterms}(r_i, r_j)$ is a number of unique GO terms occurring in the rule r_i and not occurring in the rule r_j ; the GO term a from the rule r_i is unique if it does not occur directly in the rule r_j and there is no path in GO-Inc graph that includes both term a and any term b from rule r_j premise; $\text{NoGOterms}(r)$ is the number of GO terms in the rule r premise.

4.5 Evaluation of significance of GO terms appearing in induced rules

Based on induced rule set we can also determine which of GO terms occurring in rule premises are characterized by the highest significance. A statistical significance is not the thing here but a significance (influence) of a given GO term for a quality of rule that include this term. A problem of evaluation a significance of conditions appearing in rules premises was considered in (Greco et al., 2007) where the authors applied indexes used in games theory for evaluation a quality of coalition. In conducted researches, the Banzhaf index (Banzhaf, 1965) and the modified version of the approach presented in (Greco et al., 2007) were applied to evaluate a significance of GO terms occurring in rules premises.

Let be given a rule r of the form (2). A set of GO terms occurring in the rule premise we denote by W , that is $W = \{t_1, t_2, \dots, t_n\}$. In the standard form presented in (Greco et al., 2007), the Banzhaf measure that evaluates contribution of elementary condition (a single GO term t_i here) to the rule r accuracy is calculated according to the following formula (16):

$$\phi_B(t_i, r) = \frac{1}{2^{n-1}} \sum_{Y \subseteq W - \{t_i\}} [\text{acc}(Y \cup \{t_i\}, r) - \text{acc}(Y, r)] \quad (16)$$

where $\text{acc}(Y, r)$ denotes the accuracy of r in premise part of which only GO terms included in the set Y occur; $\text{acc}(\emptyset, r) = 0$; $\text{acc}(W, r) = \text{acc}(r)$.

To evaluate a quality (significance) of a GO term in whole set of rules that create a description of analyzed gene group, it is necessary to compute its significance in each rule from the description (in each rule that include a considered GO term) and verify whether, by any chance, the GO term occurs also in rules from other gene groups descriptions. We can represent above requirements as a formula which allows evaluating GO term significance for the description of a given gene group (17):

$$G(t_i, RUL_G) = \sum_{r \in RUL_G} \phi_B(t_i, r) \text{cov}(r) - \sum_{r \notin RUL_G} \phi_B(t_i, r) \text{cov}(r) \quad (17)$$

where $\text{cov}(r)$ is the coverage (4) of the rule r , and RUL_G is the set of rules that create the description of the gene group G .

In described method of GO terms significance evaluation, contribution of each term to a rule accuracy is evaluated, while its contribution to the rule coverage is not evaluated (the coverage is considered for the whole rule only). For the purpose of a significance evaluation of GO terms describing a gene group it would be better to evaluate contribution of analyzed GO term to both accuracy and coverage. Such evaluation can be obtained by using mWS measure in formula (11) instead of the accuracy. Using mWS measure in the formula allows considering the contribution of the GO term to accuracy and coverage of all 'subrules' which can be created from the rule r . Finally, to evaluate a significance of the GO terms occurring in rule premises, a modified measure ϕ_B was applied:

$$\phi_B(t_i, r, \text{mWS}) = \frac{1}{2^{n-1}} \sum_{Y \subseteq W - \{t_i\}} [\text{mWS}(Y \cup \{t_i\}, r) - \text{mWS}(Y, r)] \quad (18)$$

where: $\text{mWS}(Y, r)$ denotes a value of the mWS measure determined for the rule r , in premise part of which only GO terms included in the Y occur; $\text{mWS}(\emptyset, r) = 0$; $\text{mWS}(W, r) = \text{mWS}(r)$.

As a consequence of the introduced modification, the measure G that evaluates quality (significance) of a GO term in whole set of rules describing analyzed gene group (19) was also modified:

$$G(t_i, RUL_G, \text{mWS}) = \sum_{r \in RUL_G} \phi_B(t_i, r, \text{mWS}(r)) - \sum_{r \notin RUL_G} \phi_B(t_i, r, \text{mWS}(r)) \quad (19)$$

Instead of mWS measure any rule quality evaluation measure can be substituted.

It is worth to mention that $\text{acc}(Y, r)$ is monotonic measure which means that if $Y \subseteq X \subseteq W$, then $\text{acc}(Y, r) \leq \text{acc}(X, r)$. It can be easily proved that, at the worst case, adding next conjunction may not improve the rule accuracy. Moreover, it is also easy to prove that mWS measure is not monotonic, because removing the condition from the premise, though can decrease accuracy of the rule, may also increase the rule coverage which can finally increase a value of mWS measure. A consequence of the reasoning is the fact that a value of the measure (18) may be negative (then (18) is the function but not the measure), while a value of the measure (16) is always nonnegative. The negative value of (18) means that 'subrules' have, at the average, better quality without the GO term t_i than 'subrules' with the term, since in such case the term t_i should be considered as unimportant (causing noises). The presented property of the measure (18) is consistent with the idea of significant descriptor, therefore lack of monotonicity is not a problem in our case.

5. Data analysis

5.1 Data description

Experiments were conducted on two freely available data sets: YEAST and HUMAN. The data set YEAST contains values of expression levels of budding yeast *Saccharomyces cerevisiae* measured in several DNA microarray experiments (Eisen et al., 1998). Our analysis were performed on 274 genes from 10 top clusters presented in the paper (Eisen et al., 1998). The data set HUMAN contains values of expression levels of human fibroblasts in response to serum (Iyer et al., 1999). In the paper (Iyer et al., 1999), 517 EST sequences were reported and divided into 10 clusters. After translation of the sequences for unique gene names and removal sequences that are duplicated or that are currently considered to be invalid, we obtained set of 368 genes. Then, each gene from YEAST and HUMAN data sets were described by GO terms from biological process ontology (BP) and from all three types of ontologies (ALL).

There were some genes in the HUMAN data set that had no GO terms assigned, thus we removed them from further analysis. After that step we obtained set consisting of 319 objects for BP ontology and 364 for all three types of ontology. To induce decision rules we created decision tables on the basis of the GO-Inc graphs for BP, CC and MF ontology. We used GO terms from at least second ontology level and describing at least three (for HUMAN data set, genes described by all three types of ontologies) or five genes from our data sets (for other data sets).

After removing from gene description GO terms that did not fulfilled the above conditions we finally obtained four decision tables. Number of genes and GO terms for each decision table are presented in Table 2.

Decision table	Number of genes	Number of GO terms
YEAST BP	274	249
YEAST ALL	274	418
HUMAN BP	310	625
HUMAN ALL	364	588

Table 2. Number of genes and GO terms obtained for each decision table.

5.2 Results of analysis

For each decision table we computed decision rules using the following criteria for Explore algorithm:

- statistical significance of the rules: 0.01,
- maximal number of descriptors in premise of the rule: 5.

The results of the analysis: number of the rules obtained (before and after filtration) and the coverage for each gene group are presented in Table 3 (for YEAST data set) and in Table 4 (for HUMAN data set).

Gene group	Decision table	Coverage	No of rules before filtration	No of rules after filtration
1	BP	100%	377	6
	ALL	100%	9625	25
2	BP	100%	1447	4
	ALL	100%	50356	14
3	BP	93%	4308	12
	ALL	100%	31708	20
4	BP	100%	43083	22
	ALL	100%	254289	50
5	BP	41%	307	4
	ALL	100%	13457	13
6	BP	93%	20225	11
	ALL	100%	4982663	25
7	BP	100%	6645	12
	ALL	100%	61361	18
8	BP	100%	3842	12
	ALL	100%	76963	18
9	BP	100%	54426	12
	ALL	100%	737049	31
10	BP	94%	810	8
	ALL	94%	124486	19

Table 3. Number of decision rules and coverage obtained for YEAST data set.

Gene group	Decision table	Coverage	No of rules before filtration	No of rules after filtration
1	BP	27%	6	6
	ALL	58%	1547	23
2	BP	53%	1204	15
	ALL	67%	16344	38
3	BP	44%	5030	11
	ALL	71%	3639	20
4	BP	88%	3937	26
	ALL	91%	87610	70
5	BP	80%	7134	8
	ALL	100%	68714	23
6	BP	90%	58263	42
	ALL	62%	84759	75
7	BP	46%	11763	8
	ALL	92%	29954	27
8	BP	26%	27	4
	ALL	37%	129	9
9	BP	43%	147	6
	ALL	86%	3875	24
10	BP	81%	16208	31
	ALL	88%	119917	48

Table 4. Number of decision rules and coverage obtained for HUMAN data set.

5.3 Interpretation of the obtained decision rules

Decision rules are generated mainly for description purposes to support drawing biological conclusions from the DNA microarray experiments. Thus, the real verification of the rule quality is its ability to provide biological interpretation of the genes composing the analyzed groups.

Below, we present two exemplary decision rules from YEAST and HUMAN data sets. We also provide their short, biological interpretation.

IF generation of precursor metabolites and energy **and**
 glucose metabolic process **and**
 hexose catabolic process **and**
 pyruvate metabolic process **and**
 hexose biosynthetic process
THEN gene group 4
 acc(r)=1.0,cov(r)=0.47, pVal=3.42e-11, FDR(r)= 1.79e-10

(20)

genes supporting the rule:
 TPI1,GPM1,PGK1,TDH3,TDH2,ENO2,TDH1,FBA1

The above rule was generated from YEAST data set, for gene group number 4, which was described as *glycolysis* in the original paper (Eisen et al., 1998). This rule is supported by 8 genes which encode proteins that are active in energy metabolism processes. The names of some of these processes can be obtained directly from GO terms that compose a conditional part of the rule. We also searched through available internet databases of publications in order to find papers that include information about all of the genes supporting the rule. One of the results of our analysis is a paper (Bruckman et al., 2007) where all genes supporting the rule are mentioned and described as the genes encoding proteins active in the glucogenesis process which is part of the glycolysis.

Another exemplary interpretation of decision rule is presented for the rule from HUMAN data set and was generated for gene group number 6.

IF protein binding **and**
 system development **and**
 cell proliferation **and**
 extracellular space **and**
 transmembrane receptor protein tyrosine kinase signaling pathway
THEN gene group 6
 acc(r)=1.0,cov(r)=0.16, pVal=0.3.3e-04, FDR(r)= 0.13

(21)

genes supporting the rule:
 FGF2,FLT1,VEGFA

The above rule is supported by three genes which encode proteins involved in the process of angiogenesis. VEGFA is one of the most important proteins from the VEGF (vascular endothelial growth factor) sub-family of growth factors. VEGFA stimulate cellular responses by binding to tyrosine kinase receptors Flt-1 and KDF/Flk-1. A VEGFA and fibroblast growth factor 2 (FGF2), both are well-investigated pro-angiogenic molecules (Kano et al.,

2005). There are evidences that VEGFA regulates expression of FLT1 (Mata-Greenwood et al., 2003) and FGF2 (Seghezzi et al., 1998).

5.4 Comparison of our method with results from Genecodis tool

We also compared results from our analysis with the results obtained from Genecodis – an Internet service for discovering co-occurrences of the annotations in the analyzed group of genes. The rule generation method provided by the authors of Genecodis is based on the Apriori algorithm for discovering frequent itemsets of attributes that describe analyzed set of objects. To generate the rules we used the following values of algorithm parameters:

- GO level: lowest level,
- Minimum number of genes: 2.

The results of a comparison: number of the rules and coverage are presented in Table 5.

Both methods, Genecodis and Explore, are based on the similar idea of searching for frequent co-occurrences of combinations of annotations, however the process of selection of attributes differs significantly in both methods. Genecodis algorithm do not perform any initial selection of the attributes that are added to the premise of created rule – it simply generates all possible combinations of GO terms. As a result of such approach one may obtain the rules that include redundant information in their premises, i.e. GO terms that are in relation $\leq$ with other GO terms composing the rule. With the Explore method, rules obtained include smaller number of GO terms, but each of the term describe different biological process.

Below we present an exemplary list of descriptors form the rule obtained from the Genecodis. This rule was generated for the objects form gene group number 6, from YEAST data set for GO terms from BP ontology:

$$\begin{array}{l} \text{IF transport and} \\ \text{ion transport and} \\ \text{ATP synthesis coupled proton transport and} \\ \text{proton transport and} \\ \text{ATP metabolic process and} \\ \text{ATP biosynthetic process and} \\ \text{THEN gene group 6} \end{array} \quad (22)$$

The analysis of the structure of GO graph for BP ontology, revealed that there are following relations among GO terms composing the above rule:

ATP synthesis coupled proton transport $\leq$ *proton transport* $\leq$ *ion transport* $\leq$ *transport*,
and

ATP synthesis coupled proton transport $\leq$ *ATP biosynthetic process* $\leq$ *ATP metabolic process*.

Gene group	Method of analysis	Coverage	Number of rules
1	Genecodis BP	45%	4
	Explore BP	100%	6
	Genecodis ALL	45%	2
	Explore ALL	100%	25
2	Genecodis BP	100%	6
	Explore BP	100%	4
	Genecodis ALL	100%	37
	Explore ALL	100%	14
3	Genecodis BP	0%	0
	Explore BP	93%	12
	Genecodis ALL	57%	5
	Explore ALL	100%	20
4	Genecodis BP	88%	23
	Explore BP	100%	22
	Genecodis ALL	100%	115
	Explore ALL	100%	50
5	Genecodis BP	0%	0
	Explore BP	41%	4
	Genecodis ALL	63%	10
	Explore ALL	100%	13
6	Genecodis BP	60%	3
	Explore BP	93%	11
	Genecodis ALL	93%	23
	Explore ALL	100%	25
7	Genecodis BP	100%	5
	Explore BP	100%	12
	Genecodis ALL	100%	5
	Explore ALL	100%	18
8	Genecodis BP	25%	2
	Explore BP	100%	12
	Genecodis ALL	92%	41
	Explore ALL	100%	18
9	Genecodis BP	80%	4
	Explore BP	100%	12
	Genecodis ALL	100%	8
	Explore ALL	100%	31
10	Genecodis BP	75%	12
	Explore BP	94%	8
	Genecodis ALL	81%	52
	Explore ALL	94%	19

Table 5. Comparison of the rules obtained from Genecodis and Explore

5.5 Significance of descriptors

The evaluation of rule descriptors significance was performed after filtration. Numbers of the rules and descriptors obtained for each gene group are presented in Table 6.

Gene group	Yeast		Human	
	Rules	$\sum$ GO terms	Rules	$\sum$ GO terms
1	6	17	6	11
2	4	12	15	20
3	12	25	11	21
4	22	31	26	40
5	4	7	8	18
6	11	11	42	63
7	12	19	8	18
8	12	22	4	8
9	12	17	6	19
10	8	12	31	56

Table 6. Number of the rules and descriptors obtained for descriptions of each gene group

Detailed results for each gene group are presented in Table 7 and Table 8, and the names of the most important and the least important GO term describing the considered group are given. Values of the significance coefficient defined by the formula (19) are also given. A ranking of the considered GO term specified by the standard method defined by formulas (16), (17) is presented in the column *Std. ranking*. Moreover, a number of descriptors describing each group are given additionally for the worst term.

As it can be noticed, there are some differences between rankings created by the standard and modified methods. We will not provide more detailed analysis of the differences as is not in the field of this paper.

The analysis of obtained results shows that establishing a ranking of GO terms that describe a given gene group provides additional knowledge about the group. Analysis of the rules and the results presented in Tables 7 and 8 leads to the conclusion that the best GO terms describing a given gene group usually occur in the best rules describing the group and do not occur in rules describing other groups (even if they occur, then they are recognized as the worst or almost the worst terms and are placed at the bottom of the other group rankings). The least significant terms usually have negative values of the coefficient (19) and are considered as the worst terms in two or more gene groups in an analyzed data set. This result is consistent with intuition and can be justified by the fact that such terms introduce a noise into gene groups description. Thus, application to the analysis some methods of postprocessing, such as rule shortening, may significantly improve the quality of obtained description. There is also interesting to verify the similarity of the groups (or at least part of genes belonging to these groups) which are described by the same, the least significant, GO terms.

Gene group	The best GO terms name (significance)	Std. ranking	The worst GO term name (significance)	Std. ranking
1	protein ubiquitination (0.78)	4	cellular macromolecule metabolic process (-0.49)	worst
2	reproductive process in single-celled organism (0.31)	5	cellular catabolic process (-0.56)	9/12
3	RNA splicing, via transesterification reactions (0.98)	2	biopolymer biosynthetic process (-1.01)	15/25
4	pyruvate metabolic process (1.53)	best	gene expression (-0.92)	worst
5	cellular respiration (1.46)	best	biopolymer biosynthetic process (-0.80)	4/7
6	purine nucleoside triphosphate biosynth. process (1.30)	best	oxidation reduction (-0.48)	worst
7	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (1.63)	4	regulation of macromolecule biosynthetic process (-1.04)	18/19
8	ribosomal small subunit biogenesis (0.98)	8	biopolymer biosynthetic process (-0.64)	worst
9	S phase (2.13)	best	biopolymer metabolic process (-0.53)	worst
10	respiratory electron transport chain (0.90)	best	generation of precursor metabolites and energy (-0.17)	11/12

Table 7. Results of evaluation of GO term significance for YEAST dataset

Gene group	The best GO terms name (significance)	Std. ranking	The worst GO term name (significance)	Std. ranking
1	cognition (0.93)	best	biopolymer metabolic process (-0.88)	worst
2	RNA biosynthetic process (0.76)	2	regulation of cellular process (-0.54)	18/20
3	innate immune response (0.59)	2	cellular alcohol metabolic process (-0.90)	worst
4	cell cycle checkpoint (1.64)	3	biopolymer metabolic process (-0.97)	39/40
5	cellular alcohol metabolic process (0.90)	best	regulation of cellular process (-1.19)	17/18
6	leukocyte migration (2.80)	best	response to stress (-0.93)	62/63
7	androgen receptor signaling pathway (1.85)	best	response to biotic stimulus (-0.49)	10/18
8	blood circulation (0.58)	best	regulation of cellular process (-1.10)	worst
9	lipid biosynthetic process (0.56)	best	regulation of cellular process (-1.29)	worst
10	skeletal system development (1.65)	best	regulation of cellular process (-1.19)	worst

Table 8. Results of evaluation of GO term significance for HUMAN dataset

6. Conclusions

Issues of gene groups description by means of GO terms were presented in the paper. Logical rules were used as a language of the description. Work connected with gene groups description by means of rules with various representations were presented. A novel method of induction, evaluation and filtration of logical rules describing gene groups was shown in the fundamental part of the paper. A method of significance evaluation of a single GO term occurring in the premises of the rules determined was also introduced. The presented method of rules induction, evaluation and filtration appeared to be very effective. After filtration we obtained small rule sets having average statistical significance better than for the unfiltered rule set.

The presented method of rules induction guarantees that all statistically significant rules are determined. Features that distinguish the proposed approach from other methods are: the way of evaluation and filtration of the rules and the fact that terms lying on the same path in Gene Ontology graph do not occur in the rule premises simultaneously (like, for example, in Genecodis system). Both, objective factors (accuracy and coverage) and subjective factors that express what is interesting for a user (premises composed of many GO terms assigned to the lowest possible level in the ontology graph) are involved in a rule evaluation. Additionally, the filtration allows the user to extract the most interesting information that can be derived from the huge number of rules.

The algorithms presented may be a useful tool that helps biologists to understand and interpret the DNA microarray experiment results. Results of the experiments presented show that the proposed method of rules induction and postprocessing is useful. In particular, the method enables to discover automatically the dependences which were found during the researches published by biologists (Bruckman et al., 2007; Mata-Greenwood et al., 2003, Kano et al., 2005; Seghezzi et al., 1998)

The algorithms described in the paper are available for the users through RuleGO Internet service (<http://rulego.polsl.pl>) (Gruca et al., 2009). The service is continuously and intensively developed. In particular, research on developing algorithms that enable rules induction in much faster way are conducted. Explore algorithm allows obtaining all statistically significant rules but, depending on the data set, calculation time may vary from several minutes to several hours, even after introducing modifications described above. Long calculation time should be regarded as the disadvantage of the presented method, thus users submitting the computational task to RuleGO service are notified by an email when their calculations are finished.

Future research will concentrate on determining the rules including descriptors referencing GO terms which does not describe the analyzed genes. In such case we will be interested in occurrence of this type of descriptors on the highest possible level of the ontology. The appropriately modified version of heuristics rules induction algorithm LEM (Grzymała-Busse & Ziarko, 2003) will be implemented in order to improve the algorithm performance.

7. References

- Agrawal, R. & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. Proc. of the 20th VLDB Conference, Santiago, Chile.
- Al-Shahrour, F.; Minguez, P.; Vaquerizas, J.M.; Conde, L. & Dopazo, J. (2005). BABELOMICS: A Suite of Web Tools for Functional Annotation and Analysis of Groups of Genes in High-Throughput Experiments. *Nucleic Acid Research*, 33, W460-W464.
- An, A. & Cercone, N. (2001). Rule quality measures for rule induction systems: Description and evaluation. *Computational Intelligence*, 17(3).
- Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M. et al. (2005). Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. *Nature genetics*, 25, 25-9.
- Baldi, P. & Hatfield G.W. (2002). *DNA Microarrays and Gene Expression*. Cambridge University Press, Cambridge.
- Banzhaf, J. F. (1965). Weighted voting doesn't work : A mathematical analysis. *Rutgers Law Review* 19, 317-343.
- Benjamini, Y. & Hochberg, T. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 59, 289-300.
- Brown, M.P.S.; Grundy, W.N.; Lin, D.; Cristianini, N.; Sugnet, C.W.; Furey, T.S.; Ares, M. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97, 262-267.
- Bruckmann, A. ; Hensbergen, P.J. ; Balog, C.I. ; Deelder, A.M. ; de Steensma, H.Y. & van Heusden G.P. (2007). Post-transcriptional control of the *saccharomyces cerevisiae* proteome by 14-3-3 proteins. *J. Proteome Res.* 6, 1689-1699.
- Bruha, I. (1997). Quality of Decision Rules: Definitions and Classification Schemes for Multiple Rules. In *Machine Learning and Statistics*, Nakhaeizadeh G., Taylor C. C. (ed.), John Wiley and Sons.
- Carmona-Sayez, P.; Chagoyen, M.; Rodriguez, A.; Trelles, O.; Carazo, J.M. & Pascual-Montano, A. (2006). Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics*, 7.
- Carmona-Saez, P.; Chagoyen, M.; Tirado, F.; Carazo, J.M. & Pascual-Montano, A. (2007). Genecodis: a web based tool for finding significant concurrent annotations in gene list. *Genome Biology*, 8.
- Cline, M.S.; Smoot M.; Cerami E. et al. (2007). Integration of biological networks and gene expression data using Cytoscape, *Nat. Protoc.*, 2, 1, 2366-82.
- Dudoit, S.; Shafer, J.P. & Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18, 71-103.
- Eisen, M.B.; Spellman, P.T.; Brown, P.O. & Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95, 14863-14868.
- Furnkranz, J. & Flach, P.A. (2005). ROC 'n' Rule Learning - Towards a Better Understanding of Covering Algorithms. *Machine Learning* 58, 39-77.
- Greco, S.; Słowiński, R. & Stefanowski, J. (2007). Evaluating importance of conditions in the set of discovered rules. (*RSFDGrC 2007*) *Lecture Notes in Artificial Intelligence* 4482, 314-321.
- Gruca, A. (2008). Analysis of GO composition of gene clusters by using multiattribute decision rules. *Biocybernetics and Biomedical Engineering*, 28, 4, 21-31.

- Gruca, A.; Sikora, M.; Chrost L. & Polanski A. (2009). RULEGO. Bioinformatical Internet Service – System Architecture. Proc. of the 16th Conference CN'09. Communications in Computer and Information Science, Wisla, Poland
- Grzymała-Busse, J.W. & Ziarko, W. (2003). Data mining based on rough sets. In: Wang J. (eds.) *Data Mining Opportunities and Challenges*. IGI Publishing, Hershey, USA, 142-173.
- Hackenberg, M. & Matthiesen, R. (2008). Annotation-modules: a tool for finding significant combinations of mulitsource annotations of gene lists. *Bioinformatics*, 24.
- Hvidstein, T. R.; Legreid, A. & Komorowski, J. (2003). Learning rule-based models of biological process form gene ontology expression time profiles using Gene Ontology. *Bioinformatics*, 19, 9, 1116-1123.
- Iyer, V.R.; Eisen, M.B.; Ross, D.T.; Schuler, Moore, G. T.; Lee, J.C.; Trent, J.M.; Staudt, L.M.; Hudson, J.; Boguski, M.S.; Lashkari, D.; Shalon, D.; Botstein, D. & Brown, P.O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science*, 283, 83-87.
- Kano, M.R. ; Morishita, Y. ; Iwata, C. ; Iwasaka, S. ; Watabe, T. ; Ouchi, Y. ; Miyazono, K. & Miyazawa, K. (2005). Vegf-a and fgf-2 synergistically promote neoangiogenesis through enhancement of endogenous pdgf-b-pdgfrbeta signaling. *J. Cell Sci.*, 118, 3759-3768, 2005.
- Khatri, P. & Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations and open problems. *Bioinformatics* 21, 3587-3595.
- Klemettinen, M. ; Mannila, H. ; Ronkainen, P. ; Toivonen, H. & Inkeri-Verkamo, A. (1994). Finding interesting rules form large sets of discovered association rules. In *The third international conference on information and knowledge management*. 29.11-02.12.1994, Gaithersburg, Maryland, USA.
- Kustra, R. & Zagdański, A. (2006). Incorporating Gene Ontology in Clustering Gene Expression Data. In *Proc. of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*.
- Maere S.; Heymans, K. & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics*, 21, 3448-3449.
- Mata-Greenwood, E.; Meyrick, B.; Soifer, S.J.; Fineman, J.R. & Black, S.M. (2003). Expression of vegf and its receptors flt-1 and flk-1/kdr is altered in lambs with increased pulmonary blood flow and pulmonary hypertension. *Am. J. Physiol. Lung Cell Mol. Physiol.*, 285, L222-L231.
- Michalski, R. S. (1983) A Theory and Methodology of Inductive Learning, In: *Machine Learnig: An Artificial Intelligence Approach*, 83-129, Tioga, Palo Alto, Calif.
- Michalski, R. S. ; Bratko, I. & Kubar, M. (1998). Machine learning and data mining. John Wiley and Sons.
- Midelfart H. (2005a). Supervised Learning in Gene Ontology Part I: A Rough Set Framework. *Transaction on Rough Sets IV LNCS*, 3700, 69-97.
- Midelfart H. (2005b). Supervised Learning in Gene Ontology Part II: A Bottom-Up Algorithm. *Transaction on Rough Sets IV LNCS*, 3700, 98-124.
- Mitchell, T. (1997). Machine learning. McGraw Hill.
- Øhrn, A.; Komorowski, J.; Skowron, A. & Synak, P. (1998). The ROSETTA software system. In *Rough Sets in Knowledge Discovery 1: Methodology and Applications*, Polkowski, L. & Skowron, A. (Ed.), 572-576. Physica-Verlag, Heidelberg, Germany.

- Pawlak, Z. (1991). *Theoretical aspects of reasoning about data*. Kluwer Academic Publisher, Dordrecht.
- Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. *Knowledge Discovery in Databases*, 229-248, Menlo Park, CA: AAAI Press.
- Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, ed. 2
- Seghezzi, G.; Patel, S.; Ren, C.J.; Gualandris, A.; Pintucci, G.; Robbins, E.S.; Shapiro, R.L.; Galloway, A.C. ; Rifkin, D.B. & Mignatti, P. (1998). Fibroblast growth factor-2 (fgf-2) induces vascular endothelial growth factor (vegf) expression in the endothelial cells of forming capillaries: an autocrine mechanism contributing to angiogenesis. *J. Cell Biol.*, 141, 1659-1673.
- Sikora, M. (2006). Rule quality measures in creation and reduction of data rule models. *Lecture Notes in Artificial Intelligence*, 4259, 716-725.
- Sikora, M. (2009). Decision rules-based data models using TRS and NetTRS – methods and algorithms. *Transaction on Rough Sets*, LNCS, Springer (in press).
- Stefanowski, J. & Vanderpooten, D. (2001). Induction of Decision Rules in Classification and Discovery-Oriented Perspectives. *International Journal of Intelligent Systems*, 16, 13-27.

IntechOpen

Engineering the Computer Science and IT

Edited by Safeeullah Soomro

ISBN 978-953-307-012-4

Hard cover, 506 pages

Publisher InTech

Published online 01, October, 2009

Published in print edition October, 2009

It has been many decades, since Computer Science has been able to achieve tremendous recognition and has been applied in various fields, mainly computer programming and software engineering. Many efforts have been taken to improve knowledge of researchers, educationists and others in the field of computer science and engineering. This book provides a further insight in this direction. It provides innovative ideas in the field of computer science and engineering with a view to face new challenges of the current and future centuries. This book comprises of 25 chapters focusing on the basic and applied research in the field of computer science and information technology. It increases knowledge in the topics such as web programming, logic programming, software debugging, real-time systems, statistical modeling, networking, program analysis, mathematical models and natural language processing.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Aleksandra Gruca and Marek Sikora (2009). Ontological Description of Gene Groups by the Multi-Attribute Statistically Significant Logical Rules, Engineering the Computer Science and IT, Safeeullah Soomro (Ed.), ISBN: 978-953-307-012-4, InTech, Available from: <http://www.intechopen.com/books/engineering-the-computer-science-and-it/ontological-description-of-gene-groups-by-the-multi-attribute-statistically-significant-logical-rule>

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2009 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the [Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License](https://creativecommons.org/licenses/by-nc-sa/3.0/), which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen